Chapter 9: Statistical Summary and Review of Quality Control Limits

Establishment of baseline pollution loads for a coal remining permit requires proper sampling and chemical analysis of pre-existing abandoned mine discharges, and the appropriate statistical analysis of flow, water quality, and pollution load data. The term "proper sampling" is taken in two contexts: (1) collection and analysis of surface water and groundwater samples, including field measurements of flow and water quality parameters, sample preservation, transportation and storage, and chemical analyses, and (2) collection of a sufficient number of samples with sampling period duration and intervals that adequately represent the variations in flow and water quality throughout the water year. Abundant scientific literature exists on collection and analytical procedures for water samples. Guidelines and protocols for water sample collection from EPA, the U.S. Geological Survey (USGS) and other sources are compiled in Table 9.1, and are discussed briefly in Chapter 1.

Table 9.1: Guidance and Protocols For Water Sample Collection

#	Type of Resource	Title	Source	HTML
1	Field Procedures	National Field Manual for the Collection of Water Quality Data	USGS	http://h2o.usgs.gov/owq/Fieldprocedures.html
2	Field Operations Manual	EMAP Surface Waters Field Operations Manual for Lakes: June, 1997 EPA/620/R-97/001	EPA	http://www.epa.gov/emjulte/html/pubs/docs/surfwatr/97fopsman.htm
3	Monitoring Guidance	Office of Water NEP Monitoring Guidance EPA-842-B-92-004	EPA	http://www.epa.gov/OWOW/estuaries/guidance/
4	Procedures	Procedures for Handling and Chemical Analysis of Sediment and Water Samples. EPA/CD-81-1	EPA/ US Army Corps of Engineers	http://www.epa.gov/owgwwtr1/info/PubList/monitoring/docs/027.pdf
5	Protocols	National Water-Quality Assessment (NAWQA) Method and Guideline Protocols	USGS	http://wwwrvares.er.usgs.gov/nawqa/protocols/doc_list.html
6	Sampling	Ground Water Sampling EPA: A Workshop Summary Nov. 30- Dec. 2, 1993. EPA/600/R-94/205	EPA	http://www.epa.gov/swerust1/cat/gwwkshop.pdf
7	Techniques	Publications on Techniques of Water Resource Investigations	USGS	http://water.usgs.gov/owq/FieldManual/chapter1/twri.html
8	Sample Preservation	Fixing Water Samples Bureau of Mines and Reclamation ID# 562-3200-203 May 1, 1997	EPA/ Bureau of Mining and Reclamation	http://www.dep.state.pa.us/dep/subject/All_Final_Techinal_guidance/bmr/562-3200-203.htm

#	Type of Resource	Title	Source	HTML
9	Sampling	Quality-control design for surface-water sampling in the National Water-Quality Assessment program (USGS Open File Report 97-223)	USGS	http://wwwrvares.er.usgs.gov/nawqa/protocols/doc_list.html
10	Sampling	Ground-Water Data-Collection Protocols and Procedures for the National Water-Quality Assessment Program: Collection and Documentation of Water-Quality Samples and Related Data (USGS Open-File Report 95-399)	USGS	http://wwwrvares.er.usgs.gov/nawqa/protocols/doc_list.html
11	Sampling	Field Guide to Collecting and processing samples of stream-water samples for the National-Water Quality Assessment program (USGS Open File Report 94-458)	USGS	http://wwwrvares.er.usgs.gov/nawqa/protocols/doc_list.html

Most of this report and EPA's *Coal Remining Statistical Support Document* (EPA-821-B-00-001) are devoted to discussion of the second context (sample period duration and interval) of proper sampling of pre-existing discharges and to the associated statistical analyses of the sample data.

The baseline pollution load is essentially a statistical summary of a data set generally consisting of 12 or more samples collected prior to issuance of a remining permit. Chapter 2 of this report provides an overview and explanation of exploratory and confirmatory statistical methods that may be used in establishing the baseline pollution load. The fundamentals of univariate, bivariate, multivariate, and time-series statistical analyses also are outlined in Chapter 2. The algorithm for analysis of mine drainage discharge data (see Figure 3.1) developed in 1987 by Dr. J.C. Griffiths and other authors of this report is described step by step in Chapter 3, and also is included in Chapter 1 of the *Coal Remining Statistical Support Document*. This algorithm was used in conducting the univariate, bivariate and time series analyses of the six relatively long term mine drainage data sets described in Chapters 4 through 8 and Appendices A through F of this report. Chapter 5 of the *Coal Remining Statistical Support Document* contains an additional 10-20 years of data on some of these six sites including data collected prior to, during, and post-remining.

The sampling plan, data collection/organization and statistical analysis components of establishing the baseline pollution load should be integrated in a continuous process. In general, abandoned mine discharges flow continuously, thus, it should not be difficult to collect an adequate number of samples. However, these discharges frequently exhibit significant variations in flow and water quality, and logistical problems may be encountered in attempting to capture

the full range and distribution of seasonal variations. Ideally, there are no missing data, and a sufficient number of samples are collected throughout the water year, at equal sampling intervals that are small enough to capture the range of natural seasonal variations. Continuous flow recorders and automated water quality samplers may be part of that ideal world, but they are rarely available or justifiable for use in remining permitting activities. Typically in routine remining permit sampling, adjustments must be made in data organization and analysis to account for missing data, unequal sampling intervals, data that are not normally distributed or that lack expression of the true extremes, and other problems.

This chapter summarizes the findings of the statistical analyses of abandoned mine discharge data contained in Chapters 4 through 8 and Appendices A through F of this report. This summary includes examples of sampling plans, data organization, univariate analysis, bivariate analysis and time series analysis, with emphasis on the practical applications of the time series. The chapter concludes with a review of the use of quality control limits for establishing and monitoring baseline pollution load at remining sites.

Sampling

The sampling plan is critical in all statistical studies and is one of the most difficult problems to resolve. One problem is the usual compromise between the samples one would like to collect and the cost of collecting them. From a research point of view, to perform a time series analysis that correctly models the variation of a parameter (e.g., flow), it is necessary to obtain observations over several years so that the model becomes truly representative. Such large collections of data are rare and the six long term data sets presented in this report are both atypical and best-case scenarios.

Another requirement that is critical for time series analysis is that the samples should be collected at equal time intervals. This criterion is almost impossible to achieve in routine sampling practice. For example, when an extreme event occurs, it is usually for at most a few days, and the common sampling intervals of one week, two weeks, or one month could easily miss the event. Secondly, if the event is a heavy snowfall or a flood, it may be physically impossible to access the sample location. The data analyzed for the studies presented in this report address these problems and other causes of unequal intervals and missing or erroneous data (e.g., loss of sample, incorrect data entry).

It is advisable to establish a sampling plan that recognizes these difficulties. It is also essential to examine the data in detail, as described in the earlier chapters of this report. It should be recognized that because of the nature of a typical data set, a rigorous statistical analysis must not be taken too far; one must compromise by being as accurate as possible without requiring impossible precision. (It is, theoretically, always possible to measure the degree of precision by replicate sampling although, in practice, replicate sampling may be too costly). The following guidelines are, therefore, a compromise and are presented as recommended guidelines only.

Sampling should be representative, cover a period of at least one year, and include both high and low flow periods within that year. Suppose 12 samples are taken at a rate of one per month for a

year. This scenario may not adequately represent baseline conditions because local extreme storm events typically occur within a few days and can result in a great range in variability between monthly samples. Extreme events are often missed with this sampling arrangement.

One recommendation for representative sample collection within the Appalachian Basin would be to use stratified sampling; divide the year into three periods of about equal length, arranged to cover high and low flow periods as follows:

January – March	April – June	September – November
high flow	intermediate flow	low flow
90 days (91 days during leap year)	91 days	91 days

The months of July, August, and December are eliminated from this recommended scenario because these months typically don't include extremes and include events covered during the other three periods. Taking one sample every 15 days within each of the three intervals would equal a total of 18 samples. Of course, to determine initial baseline pollution loading, it is preferable to increase the number of sample intervals and to extend the sampling period for more than a single year.

Data Preparation and Organization

It is always advisable to examine raw data before submitting it to analysis. The presence of unusual values and missing data usually require some kind of action. These and other features of the data set are best examined by graphical procedure. A graph of discharge or log discharge in gallons per minute versus days can be very helpful in identifying data gaps and unusual values (e.g., Figure 9.1).

Figure 9.1: Example Graph Log Discharge versus Days (Also Figure 4.2)

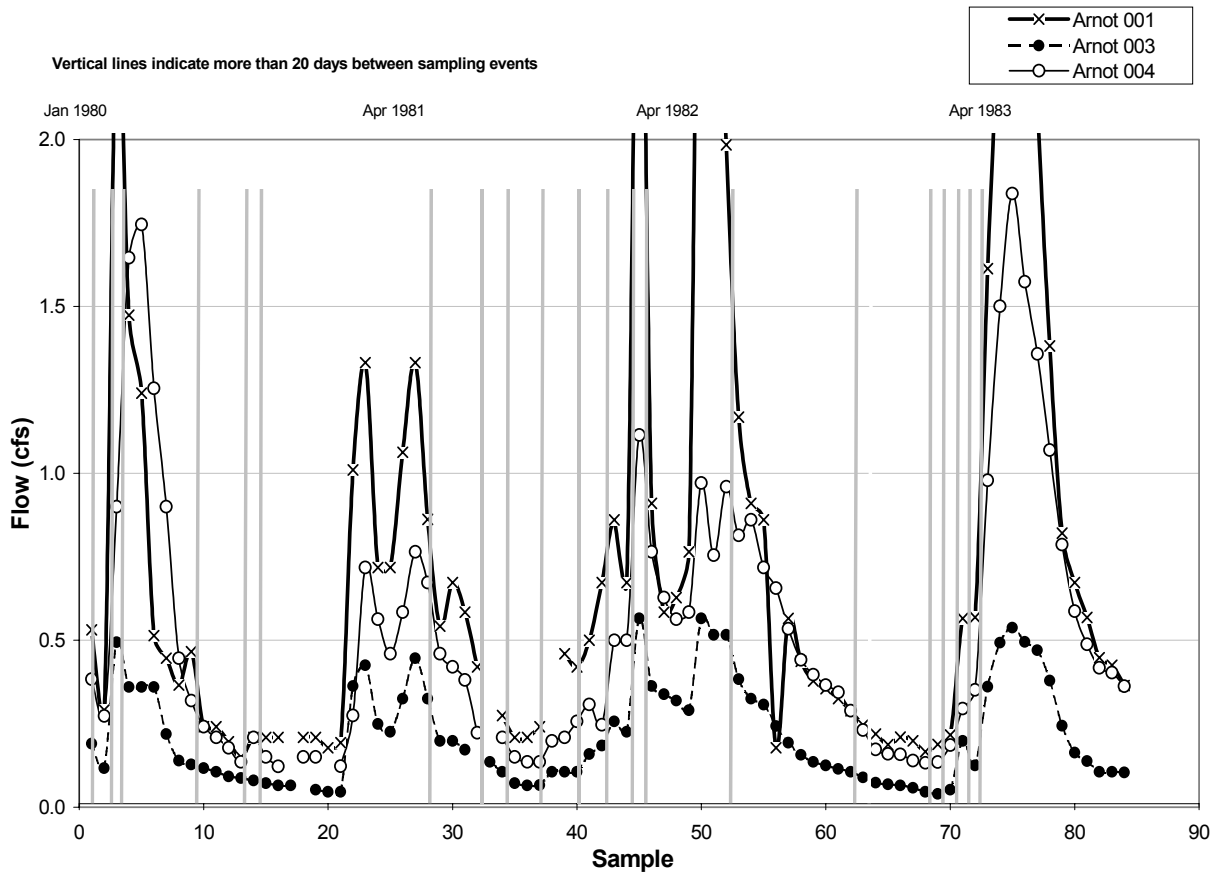


Figure 9.1 can be used to observe two kinds of information:

1) Missing values. The distribution of missing values is critical to more sophisticated analysis (particularly, time series). In general, a few missing values are not very serious, but if there are many and if they occur in clusters (Figure 9.1), the omissions may make further analysis very inexact.

   Missing values frequently occur during extreme events because during these events, sample sites are difficult to access. Sometimes, if the missing values are few and widely distributed, they may be replaced by the means (if the frequency distribution of the data is reasonably symmetrical), or by the median (if the frequency distribution of data is extremely skewed).

   In Chapters 7 and 8, a frequency distribution of the first differences between days of observation was constructed. Once constructed, both the number and the concentration density of missing observations was clearly displayed as the frequency of intervals of different lengths between observations. The variation for the Fisher site is from one day (difference = 0) to an interval of 104. The mean (26.7 days) is very nearly equal to the

median (26.5 days), thus, the distribution is roughly symmetrical around the expected sampling interval of 28 days. The central 50 percent of the distribution ($Q_1 - Q_3$) lies between 12.3 and 33 days. The most serious discrepancies are, however, that there are five observations between 70 and 104 days (four of these are 90 days or more). These large gaps in the data preclude rigorous time series analysis which requires a very close approximation to equal intervals between observations.

2) Extreme Values. The second kind of preliminary observation is to examine the data for extreme values (usually on the high side). Again, the distribution of extremes is important. Prior to examination of this data, it was believed that extreme flows would occur at regular seasonal intervals, for example, during the Spring melt. However, examination of the data presented in Figure 4.2 shows that extreme events were spread over periods from February to April (for Spring melts) and from May through June (for intense summer rains, often as thunderstorms). These wide spreads of extreme events, together with missing data (which often occurred during extreme events), made it very difficult to detect any expected true seasonal effects.

   One further point concerning extremes, is the fact that these extremes tend to introduce strong skewness (asymmetry) into the frequency distribution. This skewness is usually positive (i.e., extreme values are at the high end of the data distribution). It is conventional to apply a transformation to reduce this skewness, and logarithmic transformation is usually the most effective. It is sometimes questionable, however, to what extent the effects of extreme events should be suppressed if at all. Thus, it is prudent to examine the raw data very carefully to decide whether transformation is appropriate.

   Another effect of expressing variables in logs instead of concentration is shown in Figure 4.8, where manganese (mg/L) is plotted against log transformed discharge (cubic feet per second, cfs). There is an obvious linear association between the two variables. If discharge is expressed arithmetically in cfs (see Figure 4.9), the association is curvilinear. However, there is still a strong association between the variables. Note also that there are several outliers that appear to deviate from the trend. Expression of the log transformed data tends to suppress the effects of extreme outliers.

Univariate Analysis

The main features of the univariate statistical analyses described in Chapters 4 through 8 are the frequency distributions of the water quality parameters and flow measurement data, and the tables of summary statistics (e.g., Tables 4.1, 5.1, 6.1, 7.2, and 8.2). These tables typically include the following summary statistics: number of observations (N), number of missing observations (N*), mean, median, 10 % trimmed mean, standard deviation, standard error of the mean, minimum and maximum values (i.e., range) and quartiles. Several of these summary statistics are included in Table 1.2a of EPA's *Coal Remining Statistical Support Document* and are incorporated as conditions of remining permits (i.e., median, range, and quartiles).

An additional statistic, the coefficient of variation (CV) is included in the tables in Chapters 4-8. The coefficient of variation, usually expressed in percent (CV%), is defined as the ratio of the standard deviation to the mean multiplied by 100. This is a useful approximate guide to the degree of variation in a parameter. In general, a CV < 30% represents a stable, in control variable. In Chapters 4 through 8, most of the parameters showed much larger variation, principally because of the effects of extreme events. Use of the coefficient of variation with log transformed data may result in extreme distortion because the transformation leads to a mean of small value, resulting in a divisor of the ratio that is small and thus a CV that is inflated.

In Chapters 4 through 8, the frequency distributions of many water quality and pollution load variables were found to be normally distributed, or at least symmetrically distributed, around a value of central tendency (see for example, Figures 4.5 and 8.1e). Numerous other variables had frequency distributions that exhibited positive skewness. In Figure 5.3a, for example, there are two single observations for discharge at 50 and 80-85 gallons per minute which represent extreme events in flow. These values introduce a strong positive skewness in the histogram towards high values. In Figure 5.3b, discharge is transformed to log flow and the skewness is now towards the negative side (i.e., the transformation has over-corrected for positive skewness). In such cases, it is best not to log transform the data. Acidity (mg/L, Figure 5.4a) is somewhat symmetrical and, as would be expected, log transformation introduces a strong negative skewness (Figure 5.4b). Again, no transformation should be used.

It is possible of course, to use a less pronounced transformation (such as the square root of the variable) that may avoid the over-correction that can result from logarithmic transformation. The use of various transformations is reviewed by Tukey (1977, Chapter 3), Velleman and Hoaglin (1981, p. 46-49), and Box and Cox (1964).

Bivariate Analysis

Bivariate analysis is used to examine the relationship between pairs of variables. One expects, for example, pH and acidity or sulfate to be inversely related (as acidity increases pH declines). In the case of calcium and manganese, on the other hand, one expects positive correlation (both either increase or decrease together). The correlation coefficient (r) is used to represent the (linear) relationship between any pair of variables. The coefficient of determination (r^2), however, is a better measure of the intensity of the association between a pair of variables. For example, an r = 0.7 seems large because the range of r is from –1 to +1. However, r = 0.7 means that $r^2 = 0.49$, or that there is 49% in common between the two variables, with 51% of the variation "unexplained" by the association. For example, it would be necessary to have an r > 0.8 (i.e., > 64% in common) to claim that a strong association exists. (See Chapter 8 for additional discussion)

Another feature that can be evaluated using r and r^2 is the statistical test that accompanies a specific value of r. For example, the probability statement that for a sample size of N = 174 (see Chapter 6), a value of r > 0.124 is significantly different from zero at the 5 percent probability level, should be accompanied by the corresponding value of r^2. In Table 6.3, the correlation coefficient between pH and acidity (r = -0.365) comfortably exceeds the r (+/-) 0.124, thus, it is

statistically significant. Nevertheless, the corresponding $r^2 = 0.133$ indicates that only 13.3% of the variation is common to both variables.

Bivariate analysis of the Ernest site data also showed a strong association between all pairs of the load variables ($r^2 > 80\%$, see Figures 6.5a, b, and c). This clearly suggests that because discharge is used as a common factor in converting concentration to load, it tends to overwhelm the relationships among the other variables. This problem with pollution load variables also was detected in the analysis of data from the other sites described in Chapters 4 through 8.

In Figures 6.5a and 6.5c, the variation between the parameters increases as their values increase. This phenomenon is called heteroscedasticity and, in general, it is advisable to plot the logs of the values to make them homoscedastic. Since heteroscedastic parameters show a difference in variability with change in values, no probability statement should be made without transformation to make the variables homoscedastic. Peculiarly, the change from heteroscedasticity to homoscedasticity does not lead to a major change in the value of r. However, it does make the probability statements more reliable.

One more avenue was explored during bivariate analyses in Chapters 4 through 8, and that was to determine whether there is any lag in association between parameter pairs. The cross-correlation function is used for this purpose. The cross-correlation function calculates the linear association between observations 0 to t days apart, and thus gives an indication of when the association is strongest. The range of t is from $-\{\sqrt{N}+10\}$ to $\{\sqrt{N}+10\}$, where N is the number of observations in the series. For example, if an event occurs that affects one parameter immediately and affects another parameter five observations later, the linear correlation coefficient may be quite low at zero lag but may show a strong association after a five day lag.

Bivariate statistical analysis of data from the Fisher site (Chapter 7) can be used as an example of the use of the cross-correlation function. The correlation coefficients of zero order for each pair of variables are given in Table 7.3. The zero order value of $r = 0.663$ for acid versus iron was the highest correlation between any of the water quality parameters. The zero order correlation coefficient for iron and manganese is $r = 0.396$, and this is the maximum value. The maximum correlation coefficients and corresponding lag values from the cross-correlation functions are summarized in Table 7.4. Few are meaningful, and most are barely significant. This indicates that the degree of association was correctly represented for these variables by their conventional zero order correlation coefficients (Table 7.3).

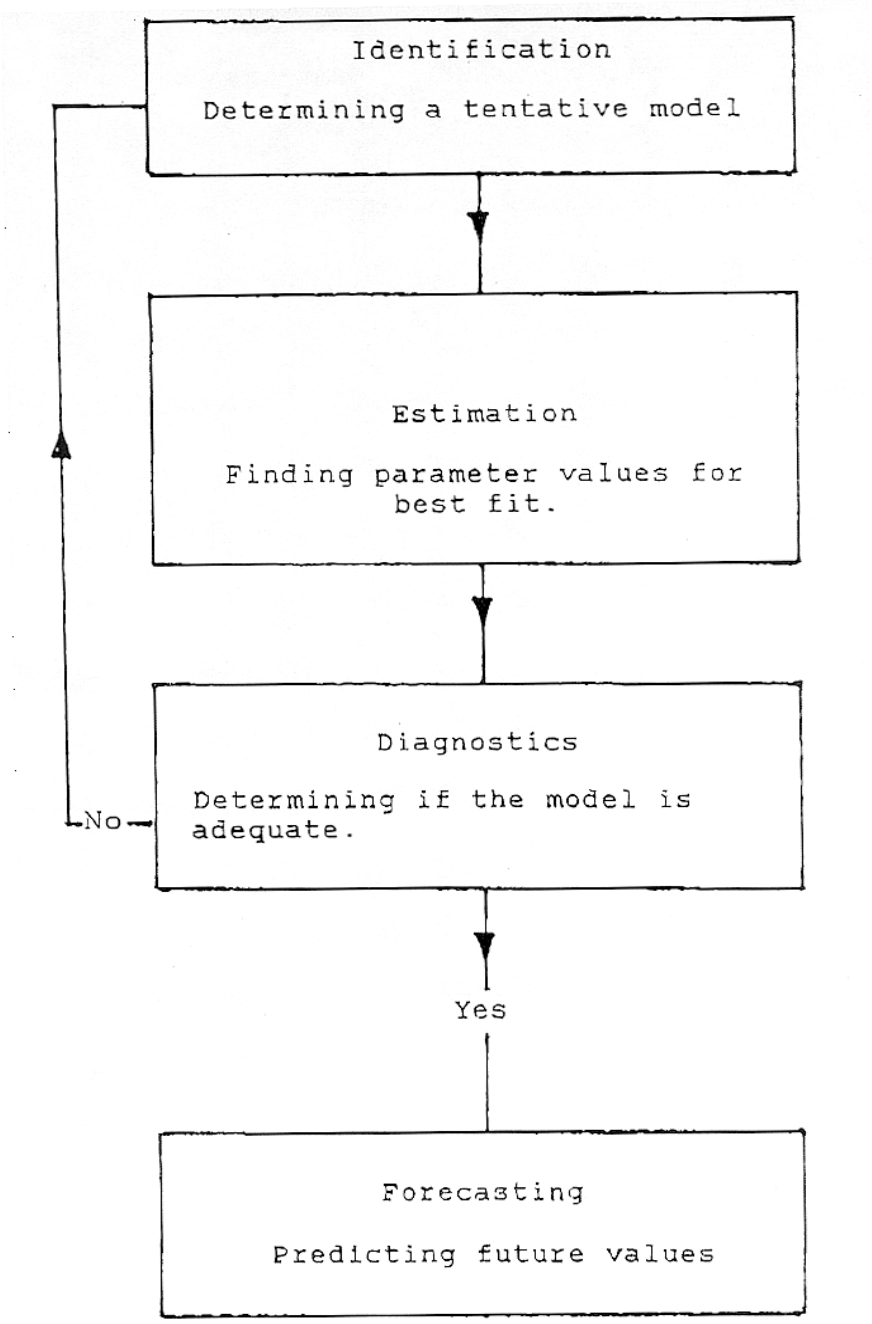
Time Series Analysis

There are two fundamental aspects to the time series analyses described in Chapters 4 through 8 and Appendix A:

1) use of a simple time series plot of the data for a particular water quality or flow variable, with or without quality control limits, to assist in evaluating patterns of variation through time (essentially an exploratory data analysis step), and

2) application of the full Box-Jenkins time series analysis and model building procedures (see Figure 9.2).

Figure 9.2: Flow Chart for Box-Jenkins Time Series Analysis



Time series analysis begins with a plot of the observations against time (days or dates). This plot is a simple outcome and can give helpful guidance to the type of time series that is represented by the variation in the data. Furthermore, the quality control limits, either some suitable multiple of (2 or 3 times) the standard deviation or, in this report, a non-parametric substitute for the standard deviation (e.g., confidence intervals around the median):

$= Md \pm 1.96 [1.25 R / (1.35 \sqrt{N'}]$
Where, R = the Interquartile Range (after McGill, R., et al., 1978).

With this application, outliers plotting beyond the confidence limits are easily seen, and the arrangement of the outliers may be either irregular (occurring as unique individuals) or systematic (e.g., periodic). Examples are given in Figure(s) 8.3.

When the plotting procedure is complete, analysis may continue using standard Box-Jenkins Time Series modeling. There is also an exhaustive Box-Jenkins procedure may be applied if there is a suitable computer package available. The main advantage of the exhaustive Box-Jenkins analysis is that very thorough testing may be performed as an automatic procedure at each stage in the analysis. The exhaustive procedure is described in many textbooks (e.g., Box and Jenkins, 1970, Nelson, 1973, and Vandaele, 1983) and the package of computer programs for pursuing the step by step analysis is also readily available in many computer systems programs (e.g., Dixon's BMDP Manuals (after 1980)).

A flow chart for Box-Jenkins time series analysis is provided in Figure 9.1. The first step is to identify a tentative model and to improve on the model by iteration through the procedure, until a more satisfactory model is found. The global model is called an ARIMA model or an Autoregressive Integrated Moving Average Model. This family of models may be summarized for convenience as an AR (autoregressive) or MA (moving average) model. A back-operator is defined as $Bz_t = z_{t-1}$ where z_t is the set of observations taken at various equally-spaced values of t (time). An autoregressive model may be represented as AR (1,0,0) which stands for an autoregressive model of order (1) with no differences (0) and no moving average terms (0); an analogous series is the MA (0,0,1). This permits extensions to AR (2), ARI (2,1,0) etc. and similarly for the MA models MA (2), IMA (0,1,2) etc. Seasonal models may be included as, for example, an ARIMA (1,1,1) (1,0,1), which represents a first order ARIMA model, together with first order seasonal autoregressive and moving average terms (Box and Jenkins, 1970, p. 322).

The basis for identification of a suitable model is the autocorrelation function (Acf) and the partial autocorrelation function (Pacf) of the observations. It is assumed that the series is stationary (i.e., the observations are free of trend). If a trend is present, it is typical to take first differences of the observations and to analyze z_{t-1} instead of z_t. In practice, it is rare to require second differences, but they are available if needed. This is where the back-operator ($Bz_t = z_{t-1}$) is useful and is why a differenced series is called integrated. The form of the Acf and Pacf is usually adequate to determine an appropriate model and one may then proceed to the estimation stage.

The variables flow, acidity and acid-load from the Ernest Refuse Pile data were chosen as examples of the use of Acf and Pacf in selecting a preliminary model. Flow shows a steady, almost straight line decline in Acf values over the first 15 lags, implying the presence of a strong trend (Figure 6.8c). This is confirmed by the corresponding Pacf which consists of a large overwhelming spike at lag 1 (Figure 6.8d). It is advisable to take first differences to remove the effect of the trend. After differencing, a first order MA (0,1,1) fits the series adequately.

The Acf for variation in log transformed acidity is entirely different in appearance, possesses at least three significant peaks at lags 1, 2, 3, and is otherwise reasonably featureless (Figure 6.8e). The corresponding Pacf shows only two spikes at lags 1 and 2 (Figure 6.8f). An MA (0,1,2) model was tried and found to be over-identified (i.e., possessed too many coefficients). For this reason an MA (0,0,2) was fitted and found adequate.

When log transformed acid-load was examined, the Acf and Pacf were almost identical to their equivalents for flow (compare Figures 6.8 c and d with 6.8 g and h). There is little doubt that flow dominates the variation when the variable is converted from concentration to load using flow as the divisor.

After complete analysis using a variety of models, it was concluded that the first order MA (0,1,1) was the most parsimonious and appropriate model for the Ernest site, and showed no significant departures from what was expected after stringent testing. The form of the equation is: $z_t = a_t - 0.247\, a_{t-1}$ with the coefficient $\hat{\theta}$ being from log acid load.

The Markson site data presented in Chapter 8 and Appendix F provides the best example of the full range of the Box-Jenkins time series analysis. The steps in the analytical procedure shown in Figure 9.1 are followed using sulfate data because it was one of the few parameters where a seasonal component appeared to be present (although never finally identified).

Identification of a tentative model was made through the Acf and Pacf of sulfate in Figures 8.4k and 8.4l. The MA (0,1,1) was chosen as a starting model because the Pacf had a single large spike (Figure 8.4l) and because this model was, in general, the most suitable for many other parameters at different sites. It was then necessary to test the residuals (i.e., the deviations of observed values from those of the fitted model). The Acf of the residuals yielded a chi-square of 41.05 with 23 degrees of freedom leading to a probability that a chi-square value as high as the one observed arising from white noise equals $0.01<P<0.02$. In other words, the chi-square is too large to be acceptable. The Acf of the residuals is summarized in Table 9.1. It can be seen that significant spikes occur at lags 3, 6, 9 (i.e., in a possible periodic recurrence usually shown by a seasonal type model). The interval of three observations in the first differences is likely to represent four in the original data, thus, the intervals are four weeks apart. If a difference of the residual is taken, the chi-square equals 145.93 with 23 degrees of freedom. Hence, the series is now over-differenced and only the differences of the initial series are required.

Table 9.2: Acf of the Residuals from Fitting an MA Model to the Original Observations After Taking a First Difference: SO_4

Lags 1-8	-0.07	0.04	0.13	0.02	-0.01	-0.12	0.00	0.01
Standard Error	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07
Lags 9-16	-0.17	-0.03	-0.12	0.10	-0.07	0.01	0.05	-0.01
Standard Error	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07

Mean = -1.972; Standard Error = 2.076; N = 252
* Spikes beyond the 16th lag unlikely to be real. Chi-square = 41.05; 0.01<P<0.02; degrees of freedom = 23

The Pacf is also listed in Table 9.2 and the significant peaks (i.e., those beyond twice their standard error) are at 3, 6, 9, and 11 or 12. This again suggests a seasonal model. It should be noted that these spikes only slightly exceed their standard error thus, the seasonal effect, if present, is weak.

Table 9.3: Pacf of the Residuals from Fitting an MA Model to the Original Observations After Taking a First Difference: SO_4

Lags 1-8	-0.07	0.03	0.14	0.04	-0.02	-0.14	-0.03	0.03
Lags 9-16	-0.13	-0.05	-0.13	0.12	-0.03	0.03	-0.01	-0.02

2 Standard Error = 2 $[1/\sqrt{252}\ ]$ = 2(0.063) = 0.126.

Examination of the residuals plotted against the date of observation shows no clear pattern of deviation. As can be seen in table 9.3, the significant residuals are arranged irregularly and occur prior to the 130th observation (out of 253 observations).

Table 9.4: Arrangement of Significant Deviations of the Residuals (> 2 Standard Error = 66.02)

Observation Numbers of Significant Residuals	
< Expected	**> Expected**
33	45
80	87 , 90
104	103
122	121

Continuing with model fitting and diagnostic testing, the next step is to examine the estimators of the parameters. For an IMA (0,1,1), there are two estimators: the coefficient of the noise term a_{t-1} ($\hat{\theta}$ = 0.725), and the overall residual standard deviation. Calculated 95% confidence limits for θ are 0.639 and 0.811, clearly confirming that the coefficient is real because the interval does not contain 0 or 1.

A number of potentially appropriate models were fitted to see if a suitable model could be found. The results of model fitting for the Markson site data are summarized at the bottom of Table 8.8). The best candidate was the IMA (0,1,1). All other models had notable failures of one or more diagnostic tests. This outcome implies that the first differences (z_{t-1}) of the original observations (z_t) represent a random walk. The seasonal effect appears to be too weak to show a positive response.

With the exception of the final step (forcasting or predicting future values), the time series examples from the Ernest and Markson sites discussed above provide a summary of the Box-Jenkins procedures listed in Figure 9.1. The last step was attempted using all the data sets presented in this report without great success. Results of this attempt using the Clarion site sulfate data are presented in Chapter 5. The reasons for this, described below, are characteristic of the six abandoned mine drainage data sets analyzed in this report.

Given the model, it is necessary to estimate the parameters for best fit. Diagnostics are applied to determine if the model is adequate and may also be used to compare different models to select the most appropriate. Finally, predictions or forecasts may be made of future values based on the selected model. This last step was shown to be of little value because extreme events inflated the confidence limits around the forecasts and thus, were not useful. There are many alternative extensions of the analytical time series procedure that can be followed, but because of extreme events, and because of difficulties with missing data and unequal values of t, it was considered imprudent to pursue the analysis further.

Quality Control Limits

The main objective of this study was to perform a statistical analysis (i.e., univariate, bivariate, and time series analyses) of numerous, long term abandoned mine drainage data sets in order to provide the foundation for developing and implementing a simple quality control approach for routine baseline pollution load analyses for remining permits. The six data sets included in Chapters 4 through 8 and Appendixes A through F of this report contain a greater number of samples (N) for a longer duration (and in some cases a tighter sampling interval) than typical remining permit baseline pollution load data sets. In addition, the statistical analyses in these chapters are more rigorous and exhaustive (see Figure 3.1) than intended for routine use in remining permits. However, much was learned from the statistical analyses of these six data sets (particularly the time series analyses) that can be applied to the use of quality control limits in establishing baseline pollution load and monitoring variations in the pollution load.

Two examples of the many variables in the six data sets are illustrated in Figures 8.3. Figure 8.3c indicates a variation in sulfate from the Markson site over a period of 253 days. Variation in total iron is shown over the same period in Figure 8.3d. As a guide to this "long range" variation, quality control limits are inserted in both graphs. One set of those limits consists of the conventional mean ($\overline{X}$) and the range between plus and minus two standard deviations (±2 $\hat{\sigma}$). The second set is calculated around the median (Md) and Tukey's recommended non-parametric limits:

$$\text{Md} \pm 1.96 [1.25 \text{ R} / \{1.35 (\sqrt{N'})\}]$$

Where, R (H-Spread) = the interquartile ($Q_3 - Q_1$) range (described in McGill et al., 1978 and Velleman and Hoaglin, 1981, p. 79).

In Figure 8.3c, the sample size for the sigma limits (N') is one. However, for the non-parametric limits, the sample size is 12. Thus, the denominator becomes 1.35 x $\sqrt{N'}$ = 1.35 x $\sqrt{12}$ = 4.677. These quality control limits are equivalent to confidence limits and are in common use in manufacturing. Quality control is maintained by choosing a sample size of, say 4, and then calculating the confidence limits around the mean of samples of size 4 which are reduced from the single sample case by the $\frac{1}{\sqrt{4}}$ = 1/2. This is, in turn, based on the relationship between the standard deviation of a single observation and the standard deviation of a mean with sample size 4 (i.e., the standard error of the $\hat{\sigma}_{\bar{x}}$ mean, as $\hat{\sigma}_{\bar{x}}$ follows (Griffiths, 1967, p.22):

$$\hat{\sigma}_{\bar{x}} = \hat{\sigma} / \sqrt{N'}$$

Where, $\hat{\sigma}_{\bar{x}}$ = the standard error of the mean.

By increasing the sample size, the confidence belts may be reduced to any desired (or affordable) limits. While this relationship holds, strictly speaking, only for a normal distribution, it is approximately true for nearly all symmetrical distributions and is substantially true for moderate departures in skewness or kurtosis.

A further series of options may be tailored to the particular problem at hand by adjusting the width of the confidence limits. In the examples discussed thus far, the limits were at the 95 percent probability level. In other words, using the range of two standard deviations, we include 95 out of every 100 observations and only 5 are expected to fall outside these limits. The same feature is approximately true of the non-parametric range. This range may be widened to 3 standard deviations, in which case about 3 observations in 1000 are expected outside the quality control limits.

In Figure 8.3c for sulfate, the two standard deviation limits emphasize the nature of the variation. Variation in sulfate starts out above the mean and above the upper confidence belt, but gradually declines with time until beyond observation number 135. Beyond observation 135, variation tends to remain within the confidence belts, and after the 230th observation, variation remains around the lower confidence belts. The quality control limits help to indicate this gradual decline despite the wide variation. The first 35 observations are persistently above the upper quality control limit, implying that some treatment of the discharge is necessary. Departures such as the 80th and 105th observations, on the other hand, are isolated events and no action is required.

The same features appear in the graph of total iron (Figure 8.3d). The earlier observations (to about 80) are mostly above the mean and around the upper quality control level. From 80 onwards, variation remains below the mean and is lowest beyond the 230th observation. In both graphs, there are some large gaps of missing observations.

In setting up baselines, and in subsequently using the baselines to judge the variation in any particular parameter, the sample size is always one so that only the conventional spread of two standard deviations and the equivalent spread measured by the interquartiles around the median are relevant. In this case the relationship: $Md \pm [1.96 \{1.25 R / (1.35 \sqrt{N'}\}]$ with $N' = 1$ reduces to $Md \pm (1.815R)$ and the calculations for sulfate and total iron are outlined in Table 8-7.

These calculations are presented to show the orders of magnitude of the different quality control limits. The rather large difference in the spreads around the mean and the median for ferrous iron (Tables 8.6 and 8.7), is essentially due to the strong negative skewness of the logs of ferrous iron. This example clearly shows that the non-parametric spread around the median is more suitable for these data. Little is lost if the distribution is symmetrical and much is gained if the data are either positively or negatively skewed.

Conclusions

The main objective of this study was to perform a statistical analysis (i.e., using univariate, bivariate, and time series approaches) of numerous, long term abandoned mine drainage data sets in order to provide the foundation for developing and implementing a simple quality control approach for routine baseline pollution load analyses for remining permits.

Sample Collection

Establishment of baseline pollution loads for a coal remining permit requires proper sampling and chemical analysis of pre-existing abandoned mine discharges, and the appropriate statistical analysis of flow, water quality, and pollution load data.

- The term proper sampling means the collection of a sufficient number of samples for a duration and at approximately constant intervals that adequately represent the variations in flow and water quality throughout the water year.

- Sampling should be representative, cover a period of at least one year, and include both high and low flow periods within that year.
- One recommendation for representative sample collection within the Appalachian Basin would be to use stratified sampling; divide the year into three periods of about equal length, arranged to cover high and low flow periods.

Discharge Variability

These pre-existing discharges frequently exhibit significant variations in flow and water quality, and logistical problems may be encountered in attempting to capture the full range and distribution of seasonal variations. There are two types of variation in pollution load that are of interest in evaluating monitoring data during and after remining to determine whether the variations are out of control compared to the established baseline conditions.

- The first and most obvious pattern of variation occurs when there are a series of extreme events, which consistently exceed the upper control level. This variation pattern indicates a sudden and dramatic increase in pollution load which may be attributed to remining, and which is referred to as the dramatic trigger.
- The second pattern of variation of concern is a trend of gradually increasing pollution load, where the general pattern of pollution load observations is increasing above the baseline central tendency value over time without exceeding the upper control level. As this second pattern of variation is much less dramatic than the first, and takes much more time and effort to detect, it is referred to as the subtle trigger. The reason that these two patterns of variation are referred to as triggers is that they can be used to initiate the requirement for a mine operator to treat a pre-existing discharge to a numeric effluent limit. If fair and reasonable consideration is given to the concerns of the mine operator and protection of the environment, the treatment triggers must be carefully established so that they are: (a) not set off prematurely or erroneously, adversely affecting the mine operator, or (b) set off too late resulting in additional mine drainage pollution without treatment.

Data Set - Initial Evaluation

The baseline pollution load is essentially a statistical summary of a data set generally consisting of 12 or more samples collected prior to issuance of a remining permit. In routine sampling for remining permits, adjustments must be made in data organization and analysis to account for missing data, unequal sampling intervals, and data that are not normally distributed or that lack expression of the true data extremes.

- It is always advisable to examine raw data before submitting it to statistical analysis. The presence of unusual values and missing data usually require some kind of action. A graph of concentration versus time or discharge or log discharge in gallons per minute versus days can be very helpful in identifying data gaps and unusual values. Missing values frequently occur during extreme events because during these events, sample sites are difficult to access.
- Another kind of preliminary evaluation is to examine the data for extreme values (usually on the high side). The wide spreads of extreme events, together with missing data (which often occur during extreme events) may make it very difficult to detect any expected true seasonal effects.

Univariate Analysis

- The main features of the univariate statistical analyses are the frequency distributions of the water quality parameters and flow measurement data, and the tables of summary statistics (e.g., Tables 4.1, 5.1, 6.2, 7.2 and 8.2).
- The frequency distribution is a graphical summary of the sample data. Its shape and accompanying summary statistics enable a greater understanding of how a parameter behaves. The normal distribution (shown in Figure 2.2) is the most widely known and most useful frequency distribution. It is also known as the bell-shaped curve.
- A major problem that is frequently encountered in the statistical analysis of water quality parameters is that the sample data are not normally distributed because it is typical to have many small valued observations in the data set and a few very large values representing extreme events. Extremes tend to introduce strong skewness (asymmetry) into the frequency distribution. This skewness is usually positive (i.e., extreme values are at the high end of the data distribution). It is conventional to apply a transformation, commonly logarithmic, to reduce this skewness (See Figure 5.3a). However, it is prudent to examine the raw data very carefully to decide whether data transformation is appropriate.
- The frequency distributions of many water quality and pollution load variables (Chapters 4 through 8) were found to be normally distributed, or at least symmetrically distributed, around a value of central tendency (see for example, Figures 4.5 and 8.1e). Numerous other variables had frequency distributions that exhibited positive skewness.
- An additional univariate statistic, the coefficient of variation (CV) is included in the Tables in Chapters 4 – 8. The coefficient of variation, usually expressed in percent (CV%), is defined as the ratio of the standard deviation to the mean multiplied by 100. This is a useful approximate guide to the degree of variation in a parameter. In general, a CV<30% represents a stable, in control variable. In Chapters 4 through 8, most of the parameters showed much larger variation, principally because of the effects of extreme events. Use of the coefficient of variation with log transformed data may result in extreme distortion because the transformation leads to a mean of small value, resulting in a divisor of the ratio that is small and thus a CV that is inflated.
- One additional parameter of interest is the number of days between sampling events. This should be approximately constant, because any outlying results could distort relationships between other parameters.

Bivariate Analysis

Bivariate analysis is used to examine the relationship between pairs of variables.

- The correlation coefficient (r) is usually used to represent the (linear) relationship between any pair of variables. The coefficient of determination (r^2) is, however, a better measure of the intensity of the association between a pair of variables. For example, r = 0.7 looks large because the range of r is from –1 to +1, but it means that $r^2 = 0.49$ or 49% of the variation is common to the two variables and therefore, 51% of the variation is "unexplained" by the association. It is necessary, therefore, to realize that one needs r > 0.8 to claim that a strong association exists; i.e., > 64% in common.

- Generally, the correlations between concentration parameters were not strong, except for those that are known to be related (e.g., pH and acidity, total and ferrous iron).
- Bivariate analysis of some data sets (e.g., Ernest site data, Chapter 6) showed a strong association between all pairs of the load variables (r^2>80%, see Figures 6.5a, b and c). This clearly suggests that because discharge is the common factor in converting concentration to load, it tends to overwhelm the relationships among the other variables.
- Heteroscedasticity occurs when the variation between the parameters increases as their values increase (see Figures 6.5a and 6.5c). In general, to correct for heteroscedasticity, it is advisable to plot the logs of the values to make them homoscedastic and to calculate correlations using log-transformed values.
- Cross-correlation analysis is performed to determine whether there is any lag in correlations between pairs of variables; i.e., to see if a relationship that is weak at zero lag is stronger at greater lags. This observation could result from a delayed effect, where one variable does not associate with another variable immediately, but only after a specific lag or period of time. For example, in a small watershed, where base flow is dominated by several large abandoned deep mine discharges, the peak of concentrations and pollution loads of acidity, iron and other parameters may occur several days or weeks following the peak of stramflow, due to the residence time in the groundwater system.
- The cross-correlation function (CCF) calculates the linear association between observation 0 to t days and so gives a picture of when the association is strongest. In the use of the cross-correlation function in bivariate and time series analyses in this report, r values of 0.2 or the more conservative r = 0.3 have been selected as critical values. This selection infers that r values less than these critical values are not significantly different than 0, and therefore can be deleted from consideration. Even if a lag correlation is significantly greater than 0, the relationship may still be weak (low r^2). In most of the examples presented in this report, there did not appear to be any very significant lag in the effects.

Time Series

Two fundamental aspects to the time series analyses (described in Chapters 4 through 8 and Appendix A) are: (1) the use of a simple time series plot of the data for a particular water quality or flow variable, with or without quality control limits, to assist in evaluating patterns of variation through time (essentially an exploratory data analysis step), and (2) the application of the full Box-Jenkins time series analysis and model building procedures (see Figure 9.2).

- Time series analysis begins with a plot of the observations against time (days or dates). This plot is a simple outcome and can give helpful guidance to the type of time series that is represented by the variation in the data. With this graph, outliers plotting beyond the confidence limits are easily seen, and the arrangement of the outliers may be either irregular (occurring as unique individuals) or systematic (e.g., periodic).
- The first step of Box-Jenkins time series analysis is to identify a tentative model and to improve on the model by iteration through the procedure, until a more satisfactory model is found. The basis for identification of a suitable model is the autocorrelation function (Acf) and the partial autocorrelation function (Pacf) of the observations. The form of the Acf and Pacf is usually adequate to determine an appropriate model and one may then proceed to the estimation stage.

- Given the model, it is necessary to estimate the parameters for best fit. Diagnostics are applied to determine if the model is adequate and may also be used to compare different models to select the most appropriate. In order to fit a model with reasonably reliable estimates, there should be at least 2-3 years of data collected at even time intervals (e.g., either weekly or monthly).
- The last step of Box-Jenkins time series analysis is to make predictions or forecasts of future values based on the selected model. This last step was shown to be of little value because extreme events inflated the confidence limits around the forecasts and thus, were not useful.
- There are many alternative extensions of the analytical time series procedure that could have been followed, but because of extreme events, and because of difficulties with missing data and unequal values of t (intervals between collection times), it was considered imprudent, for the purposes of this report, to pursue the analysis further.
- Most of the variables show the presence of a trend over time (pH, flow, acidity, acid load, iron load, ferrous iron). These variables need a first difference to remove the effects of the trend. It seems evident from the studies to date that a moving average model applied to the first differences is almost universally the best choice. In some cases, the autoregressive model, possibly with a first difference, is also appropriate. In both cases, there is an indicator that the variation in whichever parameter is being analyzed, when first differenced, leads to a random walk (the parameter is equally likely to move in one direction as the other, i.e., there is no trend).
- It is somewhat surprising that there appears to be no seasonal component in the time series models, particularly in the load variables. The only satisfactory explanation appears to be the existence of too many maxima at too many different times with very little repetition during the same time period.

Quality Control

There are many methods for defining quality control limits and there are arguments for and against all of them. Throughout this report the conventional quality control limits based upon the mean and standard deviation of the normal frequency distribution are compared to another set of non-parametric quality control limits based upon the median and other order statistics (e.g., quartiles, H-spreads, C-spreads), which may be more applicable to mine drainage data that frequently do not follow a normal distribution.

- The quality control analyses suggest that either the mean (plus or minus two standard deviations) or the non-parametric median (plus or minus a function of the H-spread) are equally appropriate. For the present, it is recommended both should be used until one or the other show superior performance.
- The quality control approach used in this report and much of statistical work in general, is dependent upon the frequency distribution of the sample data. As 95.46% of the area of the normal frequency distribution is contained in the interval of the mean +/-two standard deviations, it is expected that approximately 95 out of 100 observations will occur within these confidence intervals. In the normal frequency distribution, the values are symmetrically distributed around the mean and the mean and standard deviation are best statistical estimators of the population. In a highly skewed frequency distribution, the mean may not be

the best estimator of central tendency, and the standard deviation may not be the best measure of dispersion.

- Quality control limits can be set to compare to a specific number of remining results by setting a specific value of N′ for the equations defined in the chapters. These limits can be used as a subtle trigger for a mean or median, depending on the distribution or data. A quick trigger can also be set in the same manner by setting N′=1. For example, if one measurement is to be taken per month for a remining year, N′= 12 can be used (equation, page 3-9) to set a subtle trigger for the baseline median.
- The quality control approach should provide adjustments so that the number of monitoring samples (N) and the number of baseline samples (N) can be set to be equal when comparing these time periods (i.e., monitoring N=12 should be compared to a baseline N=12 even if the baseline contains 36 or more samples from several water years.
- Since intervals based on the median and interquartile range are non-parametric, data does not have to be transformed for normal distribution. However, it is still recommended that the data are graphed, evaluated, and transformed if transformation would improve distribution. This improved distribution would lead to improved statistical control and a tighter estimate of the confidence belts around the median.
- The analyses presented in this report were conducted using long-term data sets with frequent samples. It would be impracticial to expect this type of analysis for a remining operation. Although large data sets are preferable, the practical alternative is to employ a simple quality control approach that allows the use of data sets that are typically compiled for remining permits.